EPSE 592: Design & Analysis of Experiments

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

Repeated measures

- When you have more than one observation on the *same* sample unit, the experiment is said to contain *repeated measures*.
- Ubiquitous in the health and social sciences.
- Classic example is measuring the effect of an intervention *pre* and *post* application. In this case, average treatment effect can be quantified with a (paired) t-statistic.
- But you may want to measure the effect of an intervention at *many* points in time over the *same* sample units. This suggests an ANOVA framework.
- Another classic example: *May want patients to act as their own controls*; i.e., every sample subject receives every possible treatment (randomized, and usually with sufficient time elapsed between treatments).

Repeated measures: mathematical framework

A one-way ANOVA with repeated measures:

$$Y_{ij} = \mu_j + (s_i + \varepsilon_{ij}),$$

where Y_{ij} is the (continuous) observed response, μ_j is the *fixed* mean in the jth group of the single explanatory factor, s_i is the *random* (unique) effect of subject i, and ε_{ij} is the *random* error for subject i in group j, assumed to be independent of everything.

- Can often think of the *random effect* s_i as capturing an individual's *baseline*.
- In general, this model can be thought of as first decomposing the response Y into variability explained by the group averages (treatment effect), plus variability explained by the *individual's baseline deviation from the group average*, plus leftover/unexplained individual variation. This leads to a variance decomposition like:

$$SS(total) = SS(treatment) + SS(subjects) + SS(error)$$

Repeated measures: mathematical framework

A one-way ANOVA with repeated measures:

$$Y_{ij} = \mu_j + (s_i + \varepsilon_{ij}),$$

where Y_{ij} is the (continuous) observed response, μ_j is the *fixed* mean in the jth group of the single explanatory factor, s_i is the *random* (unique) effect of subject i, and ε_{ij} is the *random* error for subject i in group j, assumed to be independent of everything.

- A standard one-way ANOVA would not contain the random subject effect, s_i. It would thus fail to account for the *correlation between repeated measurements on the same individual*.
- It is most appropriate to treat "subject" as a *random effect* because our sample subjects are randomly selected/recruited from the target population; i.e., we are NOT, say, enlisting people that all have the same baseline response at time 1.

Repeated measures ANOVA: one-way example

Testing relative efficacy of three new pain-relieving drugs vs. placebo (so 4 treatments). Enough time is allowed between treatments so we can be sure no residual effect from previous drug. Every patient receives every treatment, with order randomized.

Subject	Placebo	Drug 1	Drug 2	Drug 3
1	5	9	6	11
2	7	12	8	9
3	11	12	10	14
4	3	8	5	8

- In a standard one-way ANOVA, each subject would receive only one treatment; in a RM-ANOVA (one-way), each subject receives all four treatments (appropriately staggered and randomized).
- Under the RM-ANOVA design, each subject acts as their own control; i.e., treatment effects are calculated as deviations between each individual treatment score and the average treatment score for each subject, thus removing between-subject baseline effects/differences.

Repeated measures ANOVA: one-way example

```
> mod <- aov(pain ~ drug + Error(subject))
> summary(mod)

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  3  70.25   23.42

Error: Within
          Df Sum Sq Mean Sq F value  Pr(>F)
drug       3  50.25  16.750   11.38 0.00204 **
Residuals  9  13.25   1.472
```

- RM-ANOVA output usually broken into *within subjects* ANOVA table and *between subjects* ANOVA table.
- Note that R refers to the "subject" effect as the "subject residuals."
- Here, we find evidence for a difference in average pain response across drug/placebo treatments (would follow up with post hocs to find exactly where).

Repeated measures ANOVA: terminology and technicalities

- The "between subject effects" refer to factors/errors that *ignore* (i.e., aggregate over) the repeated measure. These factors/errors are generally of little interest since they tell us nothing about treatment effects.
- The "within subject effects" all have something to tell us about the treatment effects, what we are usually interested in.
- The "between subjects residuals" capture what are essentially baseline differences between individuals (aggregated over all time points). Only requires estimating variance for *independent data*.
- The "within subjects residuals" capture leftover variation within each individual's responses over each time point. Requires estimating variance for *dependent data*.
- Just as our estimates of the standard error of the mean change between a *paired* vs. an *independent* samples t-test, so too do our F-tests for *within* vs. *between* subject effects.

Repeated measures ANOVA: two-way example

- Assess student confidence in math abilities after participating in two weekend workshops.
- Students complete a questionaire to assess their math confidence levels before the workshops, after the first workshop, and after the second workshop. Confidence is measured on a 20-point scale, derived from a composite score from the questionaire.
- 8 students have not taken a math course in the past 5 years (L group), 8 students have taken a math course within the past 5 years, but not within the last year (M group), and 8 students have taken a math course within the last year (H group).
- Note: Every student enters into our study with a *unique*, individual level of math confidence; hence, their baseline confidence levels are best modelled as *random*.

Repeated measures ANOVA: two-way example

- RM-ANOVA shows evidence of an marginal Treatment (time) effect, and for a differential effect of Treatment with Group (this interaction is often of greatest interest).
- Note: marginal Group effect averages responses over **all** time points, so *tells us nothing about the treatment*.

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Time	51.162	2	25.581	35.031	<.001
Time ✻ Group	18.475	4	4.619	6.325	<.001
Residual	30.670	42	0.730		

Note. Type 3 Sums of Squares

[3]

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Group	116.797	2	58.398	81.826	<.001
Residual	14.987	21	0.714		

Note. Type 3 Sums of Squares

Repeated measures ANOVA: two-way example

- RM-ANOVA shows evidence of an marginal Treatment (time) effect, and for a differential effect of Treatment with Group (this interaction is often of greatest interest).
- But how do the treatments differ in time? And do all group "trends" look different?

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Time	51.162	2	25.581	35.031	<.001
Time * Group	18.475	4	4.619	6.325	<.001
Residual	30.670	42	0.730		

Note. Type 3 Sums of Squares

[3]

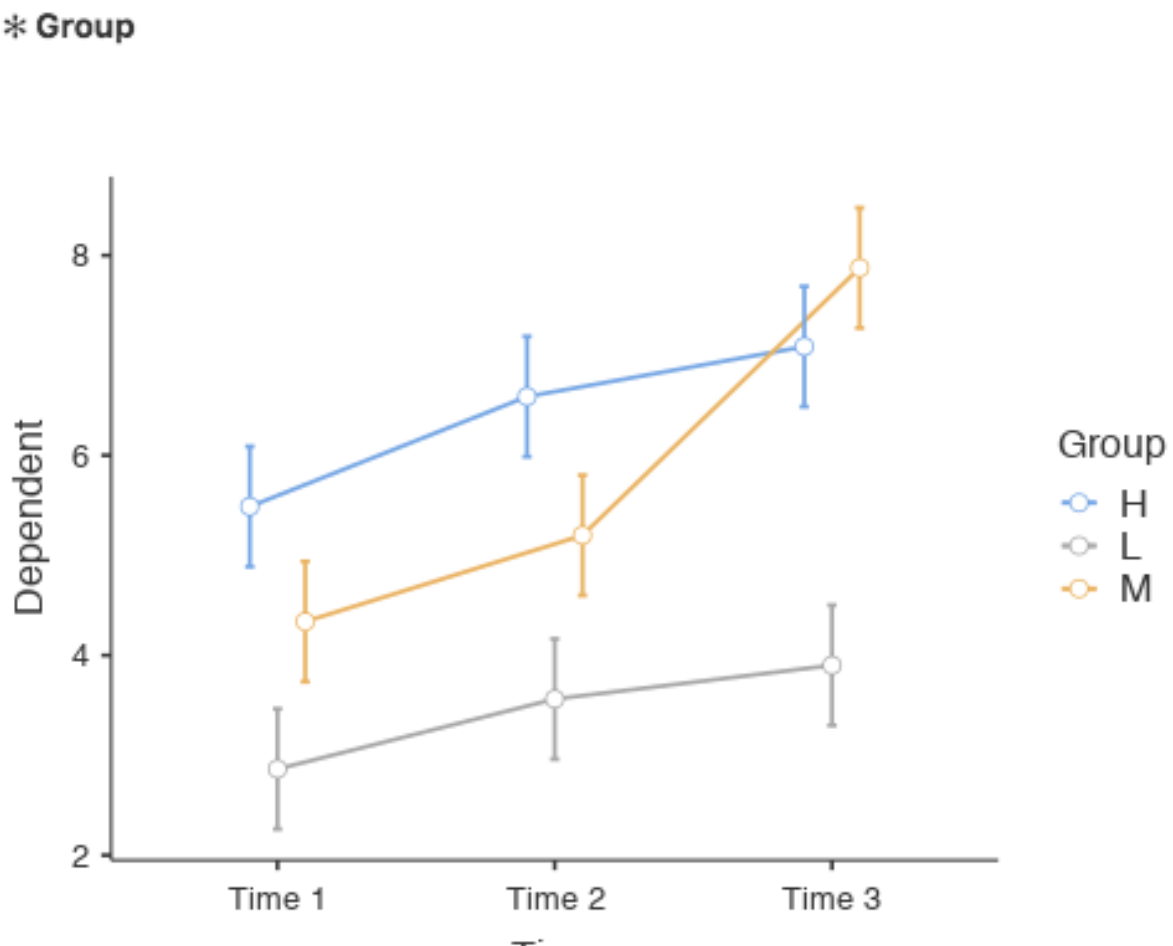
Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Group	116.797	2	58.398	81.826	<.001
Residual	14.987	21	0.714		

Note. Type 3 Sums of Squares

Repeated measures ANOVA: two-way example

- Examining interaction plot shows where differential effect(s) of treatment is(are) present. Possible explanations for differential effect?
- Also notice the obvious marginal "Group" effect expected by our design.



Repeated measures ANOVA: two-way example

- Post-hocs on marginal effect of treatment (i.e., aggregating over all three "background knowledge" groups) show strong evidence for overall intervention effect and for effect of second workshop, but only moderate evidence for effect of first workshop.
- No evidence of violation of *sphericity (compound symmetry)* assumption.

Post Hoc Comparisons - Time

Comparison			Mean Difference	SE	df	t	p_{tukey}
Time		Time					
Time 1	-	Time 2	–0.887	0.247	42.000	–3.598	0.002
	-	Time 3	–2.058	0.247	42.000	–8.344	<.001
Time 2	-	Time 3	–1.171	0.247	42.000	–4.746	<.001

Tests of Sphericity

	Mauchly's W	p	Greenhouse-Geisser ε	Huynh-Feldt ε
Time	0.933	0.501	0.937	1.000

Assumptions of repeated measures ANOVA

The assumptions for a repeated measures ANOVA are a bit different:

- Independence of observations *between* subjects/factors only (obviously, observations *within* subjects are related).
- Equality of variances (homoskedasticity) over all levels of *between* and *within* (unless more than 2 RMs) subject factors.
- Normality assumption over all levels of *between* and *within* (unless more than 2 RMs) subject factors.
- Equality of variances and normality assumption *within* factors when *more* than two repeated measurements: variances of the *differences* between all adjacent pairs of repeated measurements must be the same over all adjacent time points, and variances of the *differences* between all other possible pairs of repeated measurements must be the same over all possible pairs of time points, in addition to multivariate normality. This assumption is called *sphericity* (compound symmetry).

Checking assumptions of repeated measures ANOVA

- All the usual ANOVA diagnostics apply (e.g., examining residual plots).
- Sphericity assumption often checked by Mauchly's test and other statistics (only relevant for more than two time points).
- Sphericity is a major practical problem of implementation (when more than two time points in data). For all but the tightest controlled experiments and best behaved data, can probably count on at least some violations.
- There exist ad hoc "corrections" for the analysis when sphericity is violated (e.g., Greenhouse-Geisser, Huynh-Feldt), but better to change the model's assumed *covariance structure* if sphericity is violated.

Modelling covariance structures

- The concept of *covariance* is a generalization of the concept of *variance*.
- Recall for a random phenomenon/variable X:

$$\sigma_X^2 = \text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$$

That is, the *variance* of X is the average of the squared deviations from the mean.

- Recall: The sample analogue of the above theoretical variance is given by:

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

where $\overline{X}$ denotes the sample mean.

Modelling covariance structures

- For two random phenomena/variables X and Y, we define the *covariance* between X and Y as

$$\sigma_{XY} = \text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

- The sample analogue of the above theoretical covariance is the *sample covariance*, defined as

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

- Notice that if you plug-in X everywhere you see Y above, these two definitions reduce back to the definitions of variance and sample variance.

- Covariance is an *unstandardized* measure of correlation between X and Y; roughly, it quantifies how similarly X and Y vary.

Modelling covariance structures

- Let σ_i^2 denote the variance in the response at time point (repeated measurement) i. Let σ_{ij} denote the covariance between responses at time points (repeated measurements) i and j.
- *Independence* of observations would mean that $\sigma_{ij} = 0$ for any i and j.
- *Sphericity (compound symmetry)* requires that σ_i^2 is constant over all time points, and σ_{ij} is constant over all pairs of time points (could be different constants).
- This is a pretty unreasonable assumption much of the time (when more than 2 time points are present) because it requires that there are no changes in variation of response even in the presence of floor/ceiling effects, learning effects, accumulation effects, etc. over time.

Fundamental problems with repeated measures ANOVA

Repeated measures ANOVA has been around for a long time (100+ years); thus, the methodology is ingrained in many fields. For 2 RM designs (i.e., *pre-post* designs), RM-ANOVA is usually sufficient. However, in more generality, classical RM-ANOVA suffers from several critical flaws:

- Classical repeated measures designs do not account for sequence, carryover, learning, or other accumulation effects.
- Classical repeated measures designs do not allow for patient drop-out.
- Classical repeated measures designs require the *sphericity* assumption which is often extremely suspect in practice; moreover, RM-ANOVA is highly sensitive to violations of sphericity.

A few words on mixed effects models

- More general *linear mixed effects models* can address all the problems with RM-ANOVA (by proposing a slightly different model and set of weaker assumptions).
- General mixed effects models allow you to explicitly study, quantify, and model *dependent or confounded data* in many different ways, e.g.
  - Accumulation effects of treatment in time or space.
  - Dispersion effects of treatment in time or space.
  - Other kinds of non-stationary treatment effects in time or space.
  - Drop-out effects.
  - Nonresponse bias.
  - Measurement error.
  - Preferential sampling.
  - And much, much more!

A few words on mixed effects models

- More general RM-ANOVAs can specify different types of covariance structures than just compound symmetry.
- *Unstructured*: Every σ_i^2 and σ_{ij} is allowed to be different. Downside to this model is there are a lot of parameters to estimate, which will hurt your power and stability of your model estimates without large sample sizes.
- *Autoregressive:* Observations that are closer to each other (in time) are more correlated; e.g., $\sigma_{12} = \rho\sigma$, $\sigma_{13} = \rho^2\sigma$, $\sigma_{14} = \rho^3\sigma$.
- Can fit these models in R using the 'gls' function.
- RM-ANOVAs are special kinds of mixed effects models, and even more complex models can be specified, e.g., autoregressive moving average (ARMA) or autoregressive integrated moving average (ARIMA) models.

Practical repeated measures

- So if classical RM-ANOVA should be avoided, and we aren't learning about mixed effects modelling, then what should you do when you want to analyze repeated measures data?
- The problems with classical RM-ANOVA only really appear when we have *more than two time points* in our dataset.
- My advice: If you have more than two time points, just run *multiple* classical RM-ANOVAs on every *pair of time points* that you care about.
- Typical setup:
  - Measurements at time points 1, 2, and 3.
  - Care about possible changes in response from time point 1 to 2, and then from 2 to 3 (might also care about 1 to 3).
  - So perform two RM-ANOVAs on the two pairs of time points (1 to 2, and 2 to 3) and then adjust for the inflated Type I error rate (e.g. Bonferroni).

Repeated measures ANOVA: example

- Performing two classical RM-ANOVAs on each pair of time points yields same information as original analysis, without having to rely on the validity of the sphericity assumption.
- However, p-values not all the same (less power here).
- A bit of a multiple testing issue is present (though dependency of outcomes mitigates this concern somewhat).

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
RM Factor 1	9.452	1	9.452	17.231	<.001
RM Factor 1 ✻ Group	0.324	2	0.162	0.295	0.747
Residual	11.519	21	0.549		

Note. Type 3 Sums of Squares

[3]

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Group	64.065	2	32.033	63.856	<.001
Residual	10.534	21	0.502		

Note. Type 3 Sums of Squares

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
RM Factor 1	16.450	1	16.450	19.016	<.001
RM Factor 1 ✻ Group	13.628	2	6.814	7.877	0.003
Residual	18.167	21	0.865		

Note. Type 3 Sums of Squares

[3]

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Group	93.940	2	46.970	54.664	<.001
Residual	18.044	21	0.859		

Note. Type 3 Sums of Squares

Case study

- Case study: Bridge & Jones, 2006